You on Web

Github: https://github.com/mocho77/dsc_project

Report Submitted By
Ankur Pandey (MT18070)
Mohit Choudhary (MT18111)
Shishir Singhal (MT18106)

Table of Contents

Project Background and Motivation

The web nowadays is turning out to be the one-stop destination of information, entertainment, shopping, communication and many other utilities. A typical user nowadays uses many different web products on a frequent and daily basis. In the process, they also leave a long trail of information on these web products. If we take an example of a typical internet user who is active on the web anywhere between 10 to 20 years; then during their period on the web they may regularly visit certain websites on a frequent basis for a very long time. These websites can be social network sites like Facebook, Twitter, Instagram, etc..or video streaming sites like Youtube or knowledge sharing websites like Quora. The count of web applications where a typical user is active for a long time (10-20 years) and also use it on a frequent basis may not be very high and just maybe a maximum of 5 (just an estimated count). Our project idea is to study data collected by these applications over time and find interesting analytics which can be used to study the user's own change in interests over time.

Problem Summary

- We are active on the web from anywhere between 10-20 years.
- How many websites or applications a user visit regularly for a long time? May be maximum 5*.
- We leave a long trail of information while using these websites that we ourselves will not remember over time.
- Can we have an application that can help us view our past usage of these websites in a summarized manner?
- We aim to build a personal space of a user where he can view his usage of various web products to get interesting analytics and insights from his content consumption on these websites.
- In the current scope of this project, we are targeting a study of "Youtube Watch History" and "Android Activity" data as collected by Google and provided through Google Takeout. This can be extended to other web products gradually.

Acquiring Required Data

As our current scope was limited to Youtube and Android data, we give details below on how we acquired the required data and used it to perform our study and analysis.

One of our teammates volunteered to share his youtube watch history data and other user has volunteered to share his android activity data. We have downloaded both the data using Google Takeout as both the products are of Google. We have specifically downloaded Youtube Watch History JSON file and Android Activity and JSON file. A user can download his files and then can perform the same analysis as described in the below sections of this report on his data.

As Google only provides the video title, URLs and timestamps on which they were watched in the Watch History JSON file of Youtube dataset, we have also used Youtube API to get additional metadata about videos like duration, description, tags, categories, etc. which can make our analysis more useful.

Datasources

Youtube

- Google Takeout Youtube (https://takeout.google.com/)
- Youtube API (https://developers.google.com/youtube/v3/)
- Watch History JSON file

Android

- Google Takeout Youtube (https://takeout.google.com/)
- AppMonsta API (https://appmonsta.com/)
- Android Activity JSON and HTML file

Creating Youtube Dataset

In order to create final youtube dataset which can be used for analysis, we have followed the below steps.

- Extracted file “watch-history.json” corresponding to youtube from google takeout.
- Parsed above JSON file and collected all timestamps and video ids from video URLs in a data frame. Removed bad entries like the ones whose videos were no longer available, entries related to youtube music, etc.
- Create youtube clients and API keys for authenticated access to youtube API.
- Fetched details of videos and their channels chunk wise adhering to youtube API limits (youtube currently allows only 10k units of quota per day).
- Fetching essential details of a video/channel incurred a quota cost of 11 each, so we were able to fetch details of only 900 videos per day or 900 channels per day.
- ***This required a total of around 10 days to create the youtube dataset*** of approx. 17.5k videos and 5k Channels with 3 youtube clients.

Final Dataset: 17.5k videos with around 30 features and 5k unique channels with around 15 features.
Few attributes from video dataset: Duration, Description, Category, Tags, Audio Language, Published At, Caption, Definition, Channel Id, Statistics
Few attributes from channel dataset: Title, Country, Video Count, Keywords, Description, Published At, Statistics
Total around 8 years of data from 1st video watched in August 2011 and last in Sept 2019

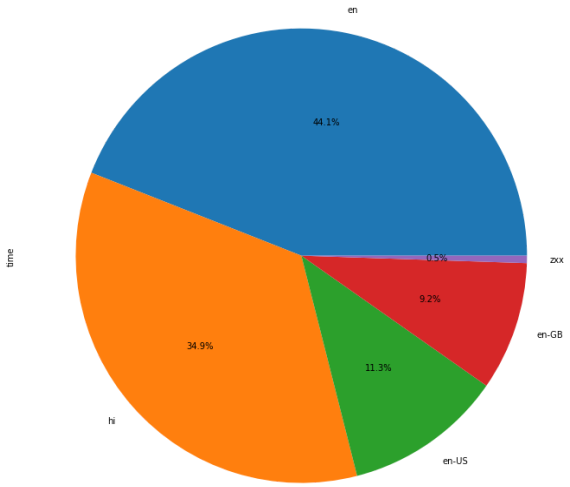
Creating Android Usage Dataset

In order to create final android dataset which can be used for analysis, we have followed the below steps.

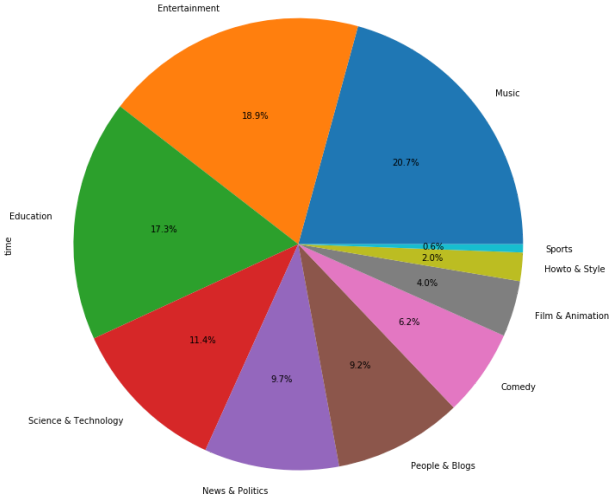
- Extracted file “android-activity.json” corresponding to Android Activity from google takeout package.
- Parsed above JSON file and collected all timestamps, app ids, app names in a dataframe.
- Fetched app category for all the unique apps using Appmonsta API.

Final dataset: 175k rows and 4 features
Dataset Attributes: Timestamp, App ID, App Name, App Category
Total around 3 years of data from 1st usage in Nov 2016 and last usage in Sept 2019

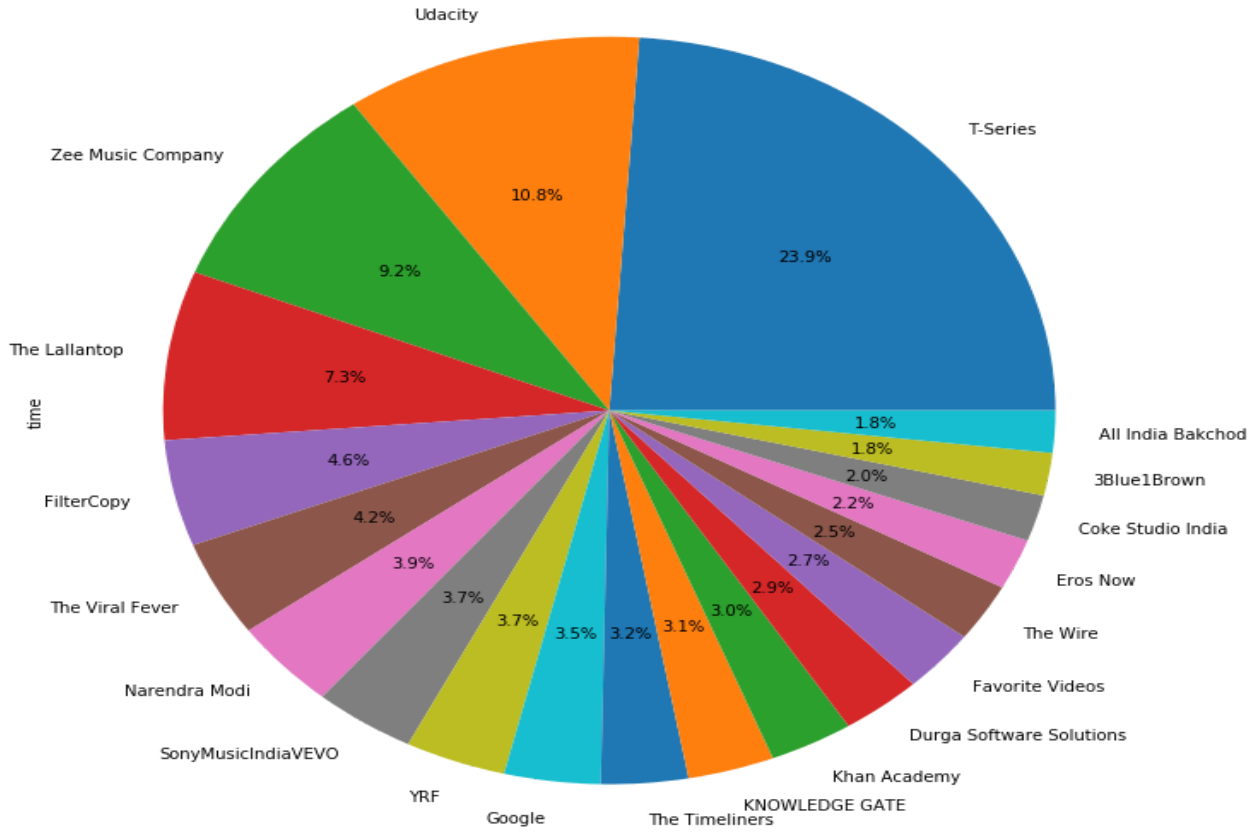
Some EDA on Youtube data



Distribution of watched videos count among languages

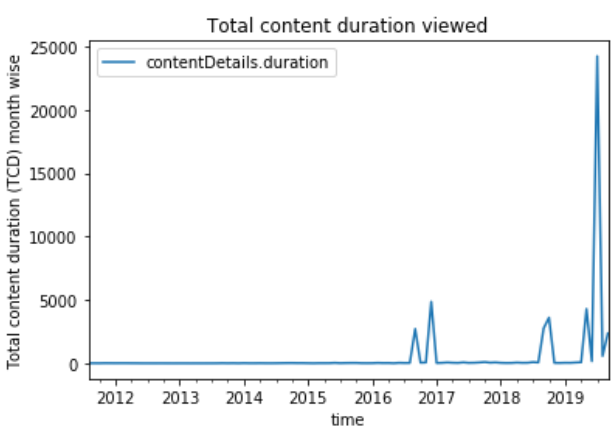


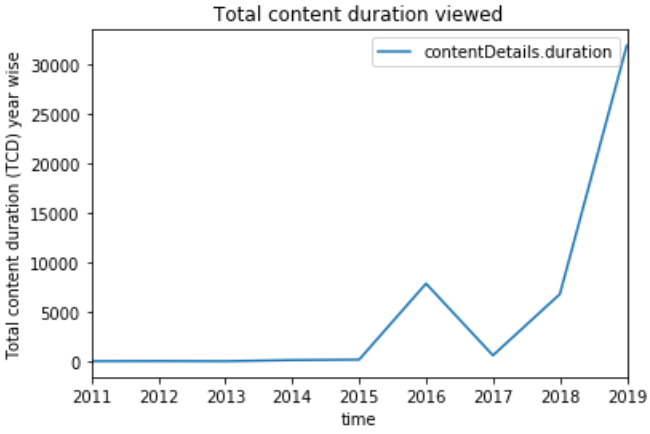
Distribution of watched videos count among categories:



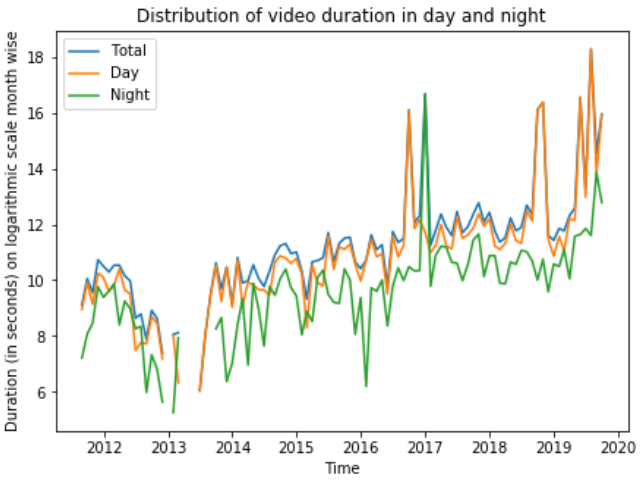
Distribution of watched videos counts among channels

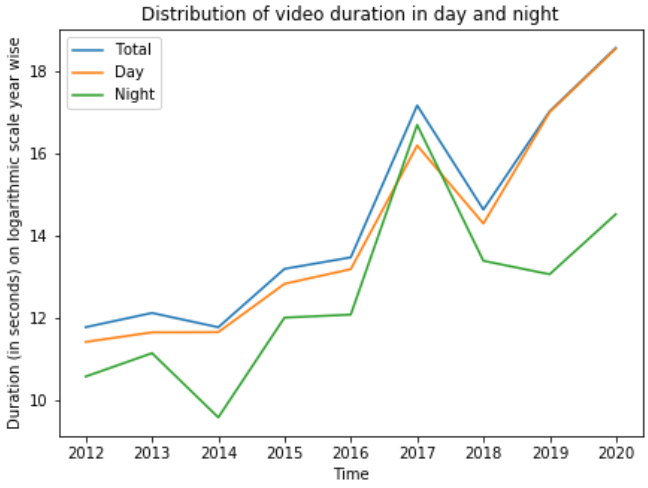
The trend in duration of watched content moth wise and year wise:

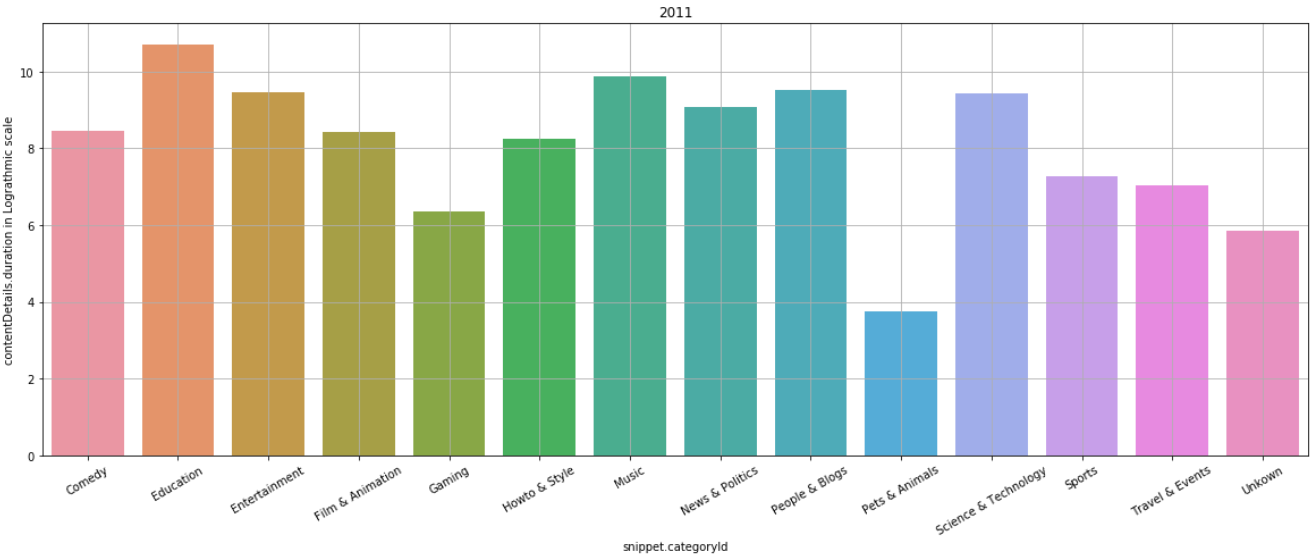




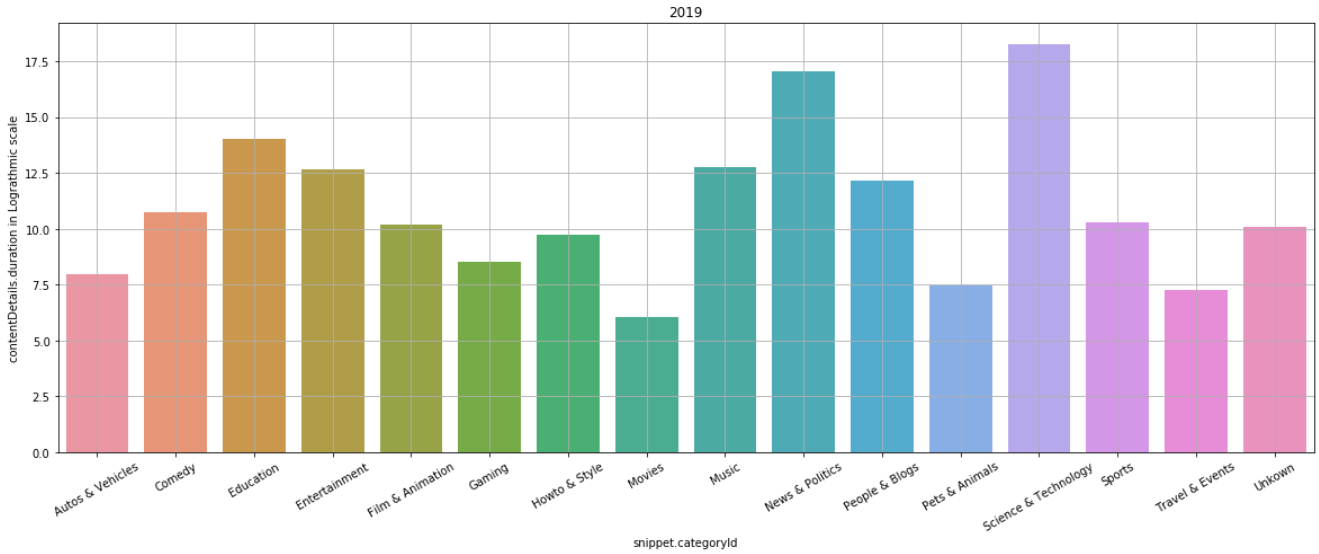
Distribution of the duration of watched content between day and night month wise and year wise:







Distribution of the duration of watched content among categories in a year-wise fashion - For the Year 2011



Distribution of the duration of watched content among categories in a year-wise fashion - For the Year 2019

Hypothesis - Text analysis

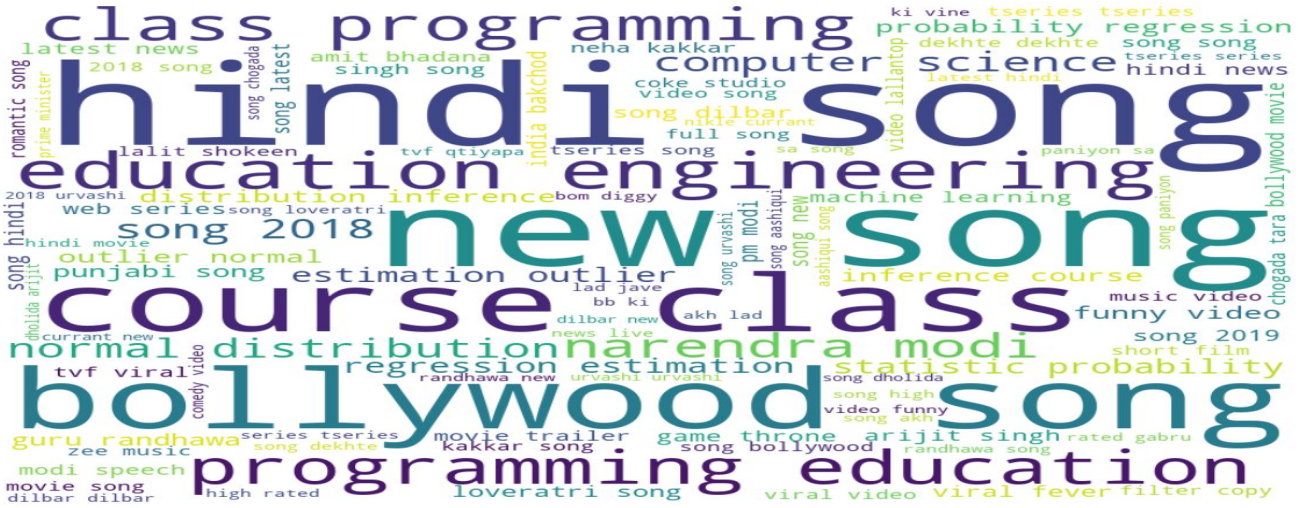
We have performed text analysis to find the most popular topics on which a user has watched videos which are not captured by the high level EDA.

The hypothesis is that if a user has spent a lot of time watching music videos then to find out what kind of artist he is listening to the most. In another example, if a user is spending time on *educational* videos, what educational topics are of most interest to the user.

We have created word clouds using TF-IDF weights which are calculated using approach described below. We reported the word cloud for bigrams. In order to obtain the word cloud of important words, we require the metric which captures the importance of the bigrams. We calculated word importance using TF-IDF values. We calculated the TF-IDF values for unigrams and bigrams. Since TF-IDF is dependent on the document and terms. In our case, the documents are details about the videos and the terms are the bigrams present in the video details. In order to obtain the TF-IDF metric independent of the video (document), we took the sum of TF-IDF of terms across all the documents.

For computing the popular words for the user, we took video's title, video's tags, video's description, video's keywords, video's channel description.

Word Cloud of Most Important words from Video title

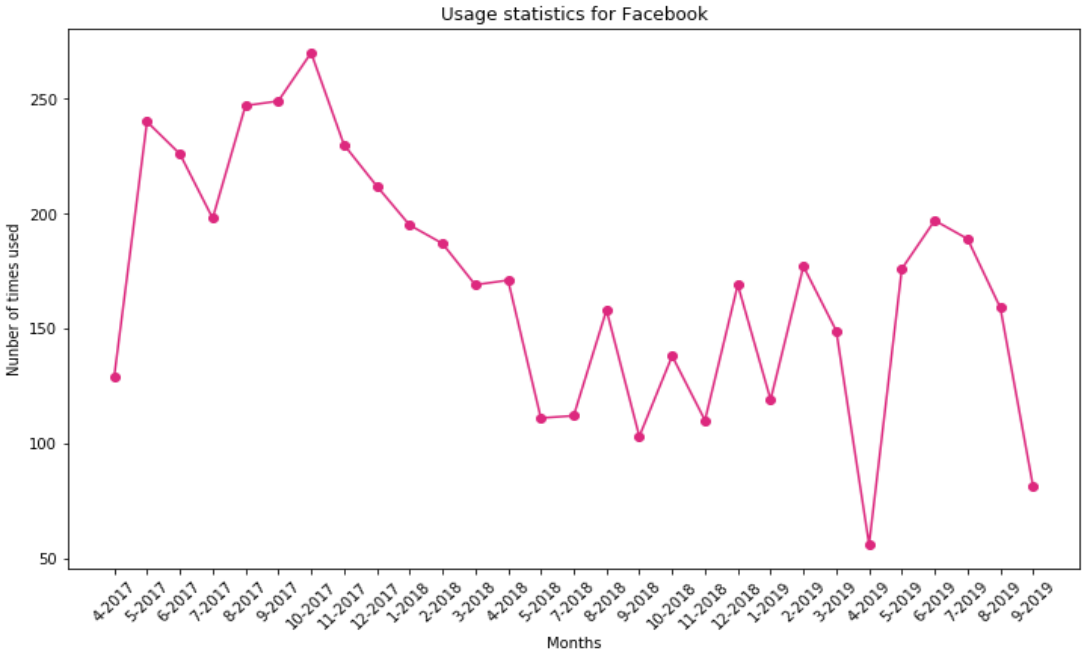


Word Cloud of Most Important words from Channel Keywords and Video tags

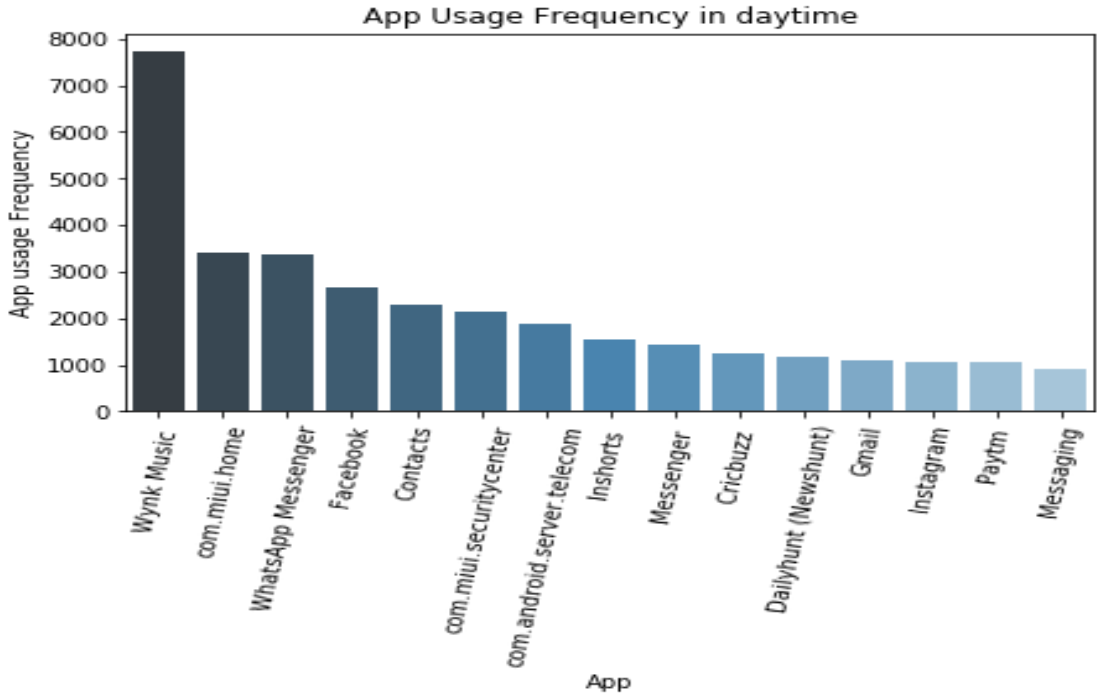
Word Cloud of Most Important words from Channel Description and Video description

Some EDA samples on Android data

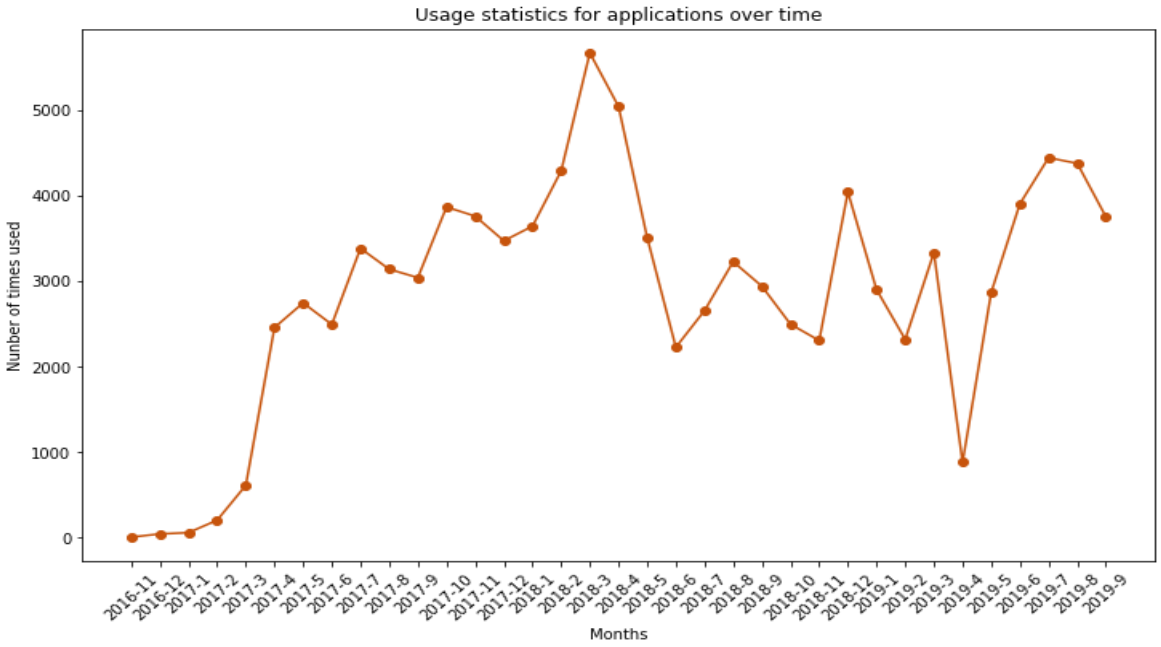
1. **Monthly application usage frequency:** Describes usage of any android application installed in your device on a monthly bases. Plot showing the usage statistics of “facebook” over time, and it is showing that facebook is highly used in the month of 9-2017.



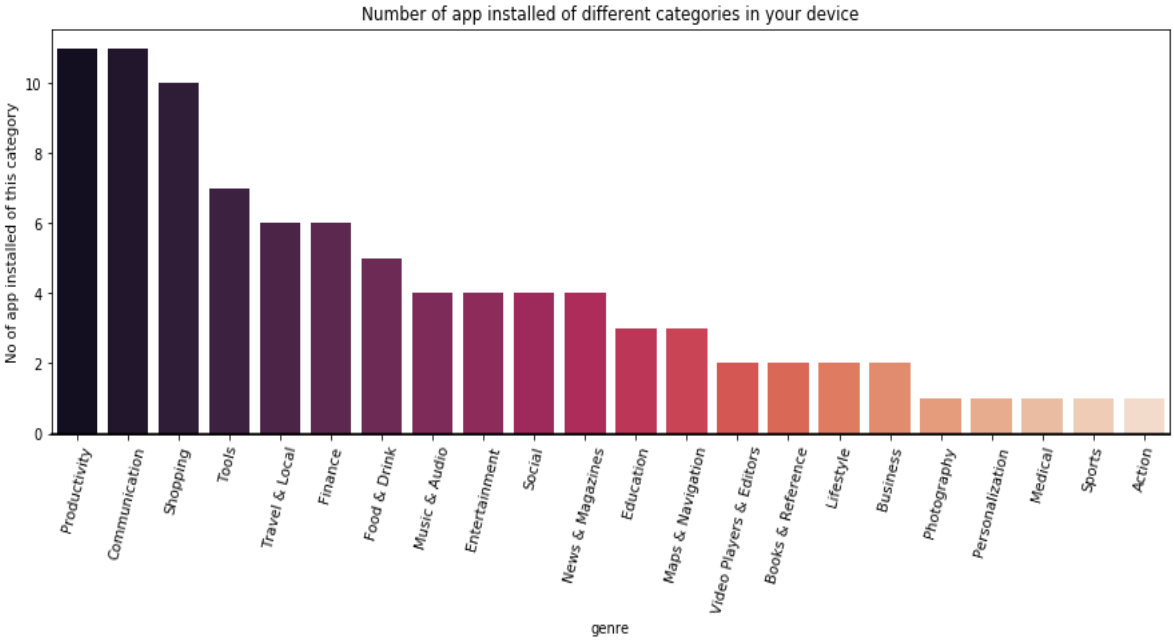
2. **Application usage frequency in daytime:** Describes which application you used most frequently in the day time. It is observed that, wynk music and whatsapp messenger is the highly used app in daytime.



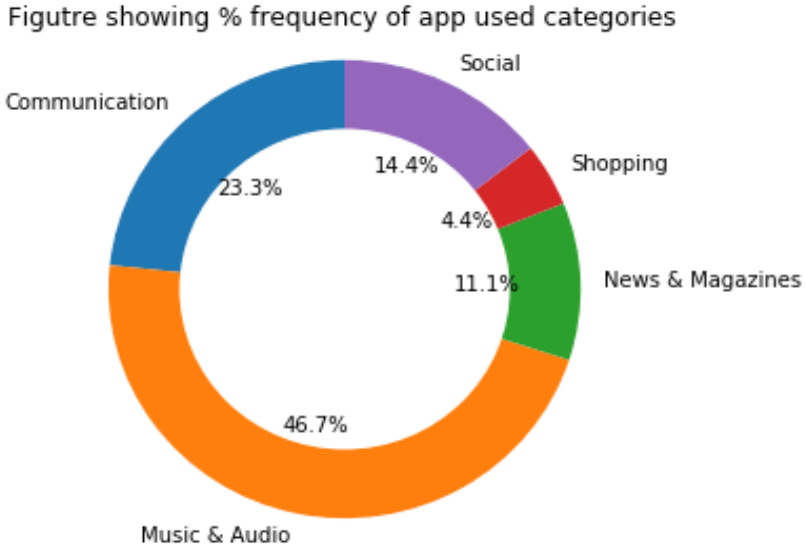
3. **Total frequency of usage of mobile applications over time:** Describes total times you have used your mobile device application in each month. It is observed that 2018-01 to 2018-03 is the peak time representing the highest usage.



4. **Number of applications installed of different category in your device:** Describing the most installed application of each category in your device, and it is observed that most of the apps belongs to the category of productivity, communication and shopping related are installed in your device.



5. **Percentage frequency of application used category-wise:** Describing the usage of each application category till now. It is observed that you are using apps related to communication and music & audio most of the time.

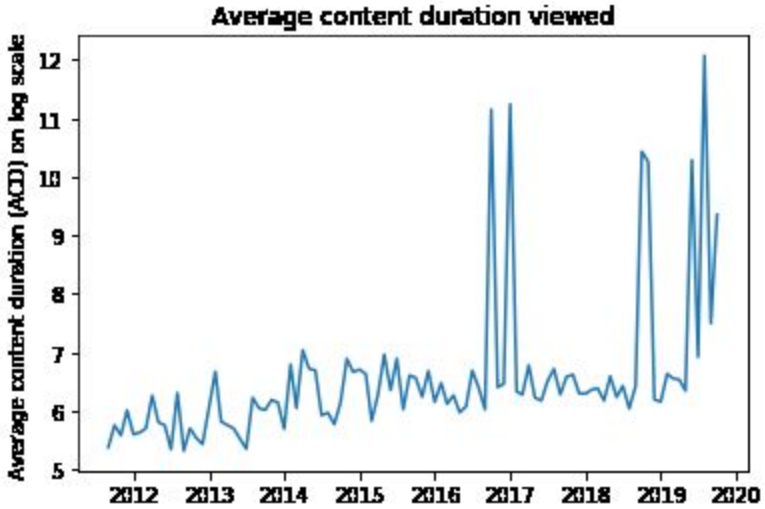


Rest all analysis and figures can be found in GitHub repository.

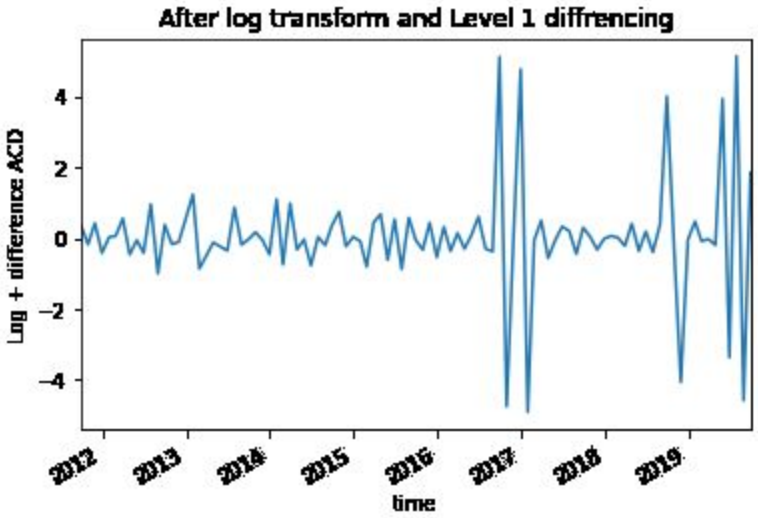
Predictive Modeling

Our project was not aimed for any predictive analysis and was focused more on analytics and visualizations for finding insights from long trial of information which is stored by various popular web products. As we have timestamps in our data, we did time series analysis which can satisfy the criteria of predictive modeling but it does not completely align with our project aim and is not really useful as a feature.

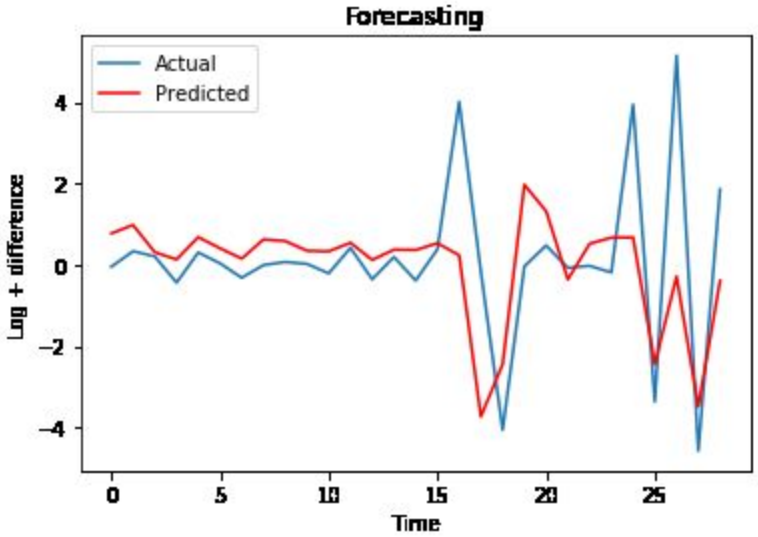
We modeled the time dependency of watched content duration for a user.



We first checked for any trend in the watched content duration and we find that there is an upward trend in the watched content duration and the same trend is also supported by the Augmented Dickey-Fuller test. The P-value for the current data is 0.11 > 0.05 which clearly greater than the critical value so we reject the null hypothesis that the watched content duration is stationary.



For time series analysis we first make the series stationary. For that, we first took the log and then performed 1 level differencing. This made the time series stationary. This was also supported by both visually as well as by the Augmented Dickey-Fuller test. We obtain the p-value of 7.193739900493205e-23 << 0.05, So we have obtained the stationary series.



Finally, we fitted the ARMA model on the resultant stationary series with hyperparameter (4, 1). The values of these hyperparameters are obtained using ACF and PACF and with manual tuning. We fitted the model on the first 70% of the time and used the trained model to forecast for the remaining 30% of the series.

After forecasting, we obtained the MSE: 2.985.

Conclusion

To conclude, after analysis of around 8 years of youtube watch history data of one of the user and 3 years of android activity data of one other user, we were able to observe many trends and useful insights. A user may forget about his usage data on these applications in the long run but the data stored by these applications can be used with our analysis and code to help users go into past and visualize the past in a summarized manner.

Reproducibility

A user can get insights on his Youtube and Android usage data as shown in the report above using the code provided in the repository (**Github:** https://github.com/mocho77/dsc_project). It also contains dataset which is used for the analysis shown above in the report.